Sun Cluster の概要 (Solaris OS 版)

Sun Microsystems, Inc.
4150 Network Circle
Santa Clara, CA 95054
U.S.A.

Part No: 819–2075–10
2005 年 8 月, Revision A

原典: Sun Cluster Overview for Solaris OS

Part No: 819-0579–10

Revision A

050815@12762

目次

はじめに

『Sun™ Cluster の概要 (Solaris OS 版)』は、Sun Cluster 製品の目的とそれを本製品でどのように達成するかを説明することによって本製品の概要を紹介するものです。本書では、Sun Cluster の主要な概念についても説明します。読者は、本書を通して Sun Cluster の特長や機能を知ることができます。

関連マニュアル

関連のある Sun Cluster のトピックについては、次の表に示したマニュアルを参照してください。Sun Cluster のマニュアルはすべて http://docs.sun.com から利用できます。

トピック	マニュアル
概要	『Sun Cluster の概要 (Solaris OS 版)』
概念	『Sun Cluster の概念 (Solaris OS 版)』
ハードウェアの設計と管理	『Sun Cluster 3.0-3.1 Hardware Administration Manual for Solaris OS』 各ハードウェア管理ガイド
ソフトウェアのインストール	『Sun Cluster ソフトウェアのインストール (Solaris OS 版)』
データサービスのインストールと管理	『Sun Cluster データサービスの計画と管理 (Solaris OS 版)』 各データサービスガイド
データサービスの開発	『Sun Cluster データサービス開発ガイド (Solaris OS 版)』
システム管理	『Sun Cluster のシステム管理 (Solaris OS 版)』

トピック	マニュアル
エラーメッセージ	『Sun Cluster Error Messages Guide for Solaris OS』
コマンドと関数のリファレンス	『Sun Cluster Reference Manual for Solaris OS』

Sun Cluster のマニュアルの完全なリストについては、お使いの Sun Cluster ソフトウェアのリリースノートを http://docs.sun.com より参照してください。

マニュアル、サポート、およびトレーニング

Sun のサービス	**URL**	内容
マニュアル	`http://jp.sun.com/documentation/`	PDF 文書および HTML 文書をダウンロードできます。
サポートおよびトレーニング	`http://jp.sun.com/supportraining/`	技術サポート、パッチのダウンロード、および Sun のトレーニングコース情報を提供します。

問い合わせについて

Sun Cluster システムのインストールや使用に関して問題がある場合は、以下の情報をご用意の上、担当のサービスプロバイダにお問い合わせください。

- 名前と電子メールアドレス (利用している場合)
- 会社名、住所、および電話番号
- システムのモデルとシリアル番号
- オペレーティング環境のリリース番号 (例: Solaris 9)
- Sun Cluster ソフトウェアのバージョン番号 (例: 3.1 8/05)

次のコマンドを使用し、システム上の各ノードに関して、サービスプロバイダに必要な情報を収集してください。

コマンド	機能
`prtconf -v`	システムメモリーのサイズと周辺デバイス情報を表示します
`psrinfo -v`	プロセッサの情報を表示する
`showrev -p`	インストールされているパッチを報告する
`prtdiag -v`	システム診断情報を表示する
`scinstall -pv`	Sun Cluster ソフトウェアのリリースおよびパッケージのバージョン情報を表示する
`scstat`	クラスタの状態のスナップショットを提供します
`scconf -p`	クラスタ構成情報を表示します
`scrgadm -p`	インストールされているリソースやリソースグループ、リソースタイプの情報を表示する

上記の情報にあわせて、`/var/adm/messages` ファイルの内容もご購入先にお知らせください。

表記上の規則

このマニュアルでは、次のような字体や記号を特別な意味を持つものとして使用します。

表 **P–1** 表記上の規則

字体または記号	意味	例
`AaBbCc123`	コマンド名、ファイル名、ディレクトリ名、画面上のコンピュータ出力、コード例を示します。	`.login` ファイルを編集します。 `ls -a` を使用してすべてのファイルを表示します。 `system%`
`AaBbCc123`	ユーザーが入力する文字を、画面上のコンピュータ出力と区別して示します。	`system%` **`su`** `password:`
AaBbCc123	変数を示します。実際に使用する特定の名前または値で置き換えます。	ファイルを削除するには、`rm` *filename* と入力します。
『 』	参照する書名を示します。	『コードマネージャー・ユーザーズガイド』を参照してください。

表 **P–1** 表記上の規則　*(続き)*

字体または記号	意味	例
「 」	参照する章、節、ボタンやメニュー名、強調する単語を示します。	第 5 章「衝突の回避」を参照してください。 この操作ができるのは、「スーパーユーザー」だけです。
\	枠で囲まれたコード例で、テキストがページ行幅を超える場合に、継続を示します。	`sun% grep '^#define \` `XV_VERSION_STRING'`

コード例は次のように表示されます。

- C シェル

  `machine_name% command y|n [filename]`

- C シェルのスーパーユーザー

  `machine_name# command y|n [filename]`

- Bourne シェルおよび Korn シェル

  `$ command y|n [filename]`

- Bourne シェルおよび Korn シェルのスーパーユーザー

  `# command y|n [filename]`

[ ] は省略可能な項目を示します。上記の例は、*filename* は省略してもよいことを示しています。

| は区切り文字 (セパレータ) です。この文字で分割されている引数のうち 1 つだけを指定します。

キーボードのキー名は英文で、頭文字を大文字で示します (例: Shift キーを押します)。ただし、キーボードによっては Enter キーが Return キーの動作をします。

ダッシュ (-) は 2 つのキーを同時に押すことを示します。たとえば、Ctrl-D は Control キーを押したまま D キーを押すことを意味します。

第 1 章

Sun Clusterの概要

SunPlex システムはハードウェアと Sun Cluster ソフトウェアが統合されたソリューションであり、高度な可用性とスケーラビリティーを備えたサービスを提供するために使用されます。 この章では、Sun Cluster 機能の概要を説明します。

この章には、以下の節があります。

Sun Cluster によるアプリケーションの可用性の向上

クラスタとは、単一のシステムとして連係して動作する 2 つ以上のシステムまたはノードのことで、アプリケーションやシステムリソース、データをユーザーに提供する連続的な可用性を備えたシステムです。クラスタの各ノードは、それぞれが十分に機能するスタンドアロンシステムです。しかし、クラスタ環境では、すべてのノードがインターコネクトによって接続され、単一のエンティティーとして動作しますので、可用性と性能が向上します。

HA を備えたクラスタは、通常、単一のサーバーシステムなら停止するような障害が発生しても、データやアプリケーションに対してほとんど連続的なアクセスを提供するように稼動し続けることができます。ハードウェア、ソフトウェア、またはネットワークの単一の故障によりクラスタに障害が発生することはありません。これに対して、フォルトトレラントのハードウェアシステムは、データとアプリケーションに対する一定したアクセスを可能にしますが、特殊なハードウェアが必要なため、コストが高くなります。フォルトトレラントシステムには通常、ソフトウェア障害に対する備えはありません。

個々の Sun Cluster システムは密接に関わり合ったノードの集合であり、すべてのネットワークサービスやアプリケーションが一元的に管理されます。Sun Cluster システムは、次のハードウェアとソフトウェアの組み合わせを通して HA を実現します。

- 冗長化されたディスクシステム。ストレージを提供するこれらのディスクシステムは一般にミラー化されるため、ディスクやサブシステムに障害が発生しても、操作が中断されることはありません。さらに、ディスクシステムへの接続は冗長化されているため、サーバーやコントローラ、ケーブルに障害が発生しても、データにアクセスできなくなることはありません。ディスクに直結されていないノードからリソースへのアクセスは、ノード間を結ぶ高速インターコネクトを通して行われます。さらに、クラスタのすべてのノードがパブリックネットワークに接続されているため、複数のネットワークに散在するクライアントからクラスタにアクセスできます。
- 電源装置や冷却システムなど、冗長化されたホットスワップ可能コンポーネント。これらのコンポーネントは冗長化されているため、ハードウェアに障害が発生しても、システムは操作を続けることができ、可用性が向上します。ハードウェアコンポーネントがホットスワップ可能であれば、そのコンポーネントを動作中のシステムから取り外したり、システムに追加することができます。そのためにシステムを停止する必要はありません。
- Sun Cluster ソフトウェアフレームワーク。このフレームワークはノードの障害を素早く検知し、それと同一環境で動作する別のノードにアプリケーションやサービスを移行します。すべてのアプリケーションが同時に使用不能になることはありません。停止したノードと関係のないアプリケーションは、この復旧処理の間も全面的に使用可能です。さらに、障害が発生したノードのアプリケーションは、復旧されると同時に使用可能になります。復旧したアプリケーションは、ほかのすべてのアプリケーションが完全に復旧するまで待つ必要はありません。

可用性の管理

システムで単一ソフトウェアまたはハードウェアの障害が発生してもあるアプリケーションが稼働し続けられる場合、そのアプリケーションには高い可用性があります。ただし、アプリケーション自体のバグやデータは破損に起因する障害の場合は除きます。HA のアプリケーションには次が適用されます。

- リソースを使用するアプリケーションから、復旧は透過的に行われます。
- リソースのアクセスは、ノードに障害が発生しても完全に保持されます。
- アプリケーションのホストノードが別のノードに移行されたことをアプリケーションが検知することはありません。
- 単一ノードの障害は、このノードに接続されているファイルやデバイス、ディスクボリュームを使用する障害を受けないほかのノード上のプログラムへ、完全に透過的に行なわれます。

フェイルオーバーサービスとスケーラブルサービス、およびパラレルアプリケーション

フェイルオーバーサービスやスケーラブルサービス、パラレルアプリケーションを使用すると、アプリケーションの高い可用性が実現し、クラスタで動作するアプリケーションの性能が向上します。

フェイルオーバーサービスでは、冗長性を通して HA を提供します。障害が発生した場合、ユーザーが介入することなく、アプリケーションの設定に従って、稼動しているアプリケーションを同じノードで再起動するか、クラスタの別のノードに移動することができます。

スケーラブルサービスでは、性能を高めるために、クラスタの複数のノードでアプリケーションを同時に実行します。スケーラブルな構成では、クラスタ内の各ノードが、データを提供して、クライアント要求を処理することができます。

PDB (パラレルデータベース) を使用すれば、データベースサーバーの複数のインスタンスを使って次のことができます。

- クラスタに参加する。
- 同じデータベースに対する別々のクエリーを同時に処理する。
- 大規模なクエリーの場合、クエリーを並列に処理する。

フェイルオーバーサービスとスケーラブルサービス、およびパラレルアプリケーションの詳細については、26 ページの「データサービスのタイプ」を参照してください。

IP ネットワークマルチパス

クライアントは、パブリックネットワークを介してクラスタにデータ要求を行います。各クラスタノードは、1 つまたは複数のパブリックネットワークアダプタを介して少なくとも 1 つのパブリックネットワークに接続されています。

IP ネットワークマルチパス では、サーバーの複数のネットワークポートを同じサブネットに接続できます。IP ネットワークマルチパス ソフトウェアはネットワークアダプタ障害からの復旧をサポートします。そのために、まず、ネットワークアダプタの障害や修復を検知し、次に、アダプタと代替アダプタとの間でネットワークアドレスを同時に切り替えます。複数のネットワークアダプタが機能している場合、IP ネットワークマルチパス は、送信パケットをアダプタ間に分配することによってデータスループットの向上を図ります。

記憶装置の管理

多重ホストストレージではディスクが複数のノードに接続されるため、ディスクの高い可用性が実現します。この場合、データには複数のパスを通してアクセスできるため、1 つのパスに障害が発生しても、別のノードがその代わりをします。

多重ホストディスクの使用によって、次のクラスタ処理が可能になります。

- 1 つのノードに障害が発生しても動作を続ける。
- アプリケーションデータやアプリケーションバイナリ、構成ファイルを一元化する。
- ノードの障害からユーザーを保護する。クライアント要求があるノードを介してデータにアクセスしていて失敗した場合、これらの要求は、同じディスクへの直接接続を持つ別のノードを使用するようにスイッチオーバーされます。
- ディスクを「マスター」する稼動系を通して広域にアクセスするか、ローカルパスを通して直接かつ並列にアクセスする。

ボリューム管理のサポート

ボリュームマネージャーを使用すると、大量のディスクやそこに格納されているデータを管理することができます。ボリュームマネージャーは、次のような機能を使ってストレージの容量やデータの可用性を高めます。

- ディスクドライブのストライピングやコンカチネーション
- ディスクのミラー化
- ディスクドライブのホットスペア
- ディスク障害への対応とディスクの交換

Sun Cluster システムは、次のボリュームマネージャーをサポートします。

- Solaris ボリュームマネージャー
- VERITAS Volume Manager

Sun StorEdge Traffic Manager

Sun StorEdge Traffic Manager ソフトウェアは、Solaris オペレーティングシステム 8 からそのコア入出力フレームワークに完全に組み込まれています。Sun StorEdge Traffic Manager ソフトウェアを使用すると、Solaris オペレーティング環境の単一インスタンス内で複数の入出力コントローラインタフェースを通してアクセスされるデバイスの表現や管理が効果的になります。Sun StorEdge Traffic Manager アーキテクチャーには、次の機能が備わっています。

- 入出力コントローラの障害による入出力の中断を防止する。
- 入出力コントローラの障害時に代替のコントローラに自動的に切り替える。
- 複数の入出力チャネルに負荷をロードバランスさせることによって、入出力の性能を高める。

ハードウェア RAID (redundant array of independent disks) サポート

Sun Cluster システムでは、ハードウェア RAID (Redundant Array of Independent Disks) やホストベースのソフトウェア RAID が使用できます。ハードウェア RAID では、ストレージアレイまたはストレージシステムのハードウェアの冗長性を使って、

個々のハードウェア障害がデータの可用性に影響がないようにします。別々のストレージアレイ間でデータがミラー化されている場合には、ホストベースの RAID を使って、個別のハードウェア障害 (ある1つのストレージアレイが完全にオフライン) がデータの可用性に影響がないようにします。ハードウェア RAID とホストベースのソフトウェア RAID を同時に使用することもできますが、ある程度の高いデータ可用性を維持するために、1 つの RAID ソリューションだけを使用することもできます。

ファイルシステムのサポート

クラスタシステム本来の特性の 1 つにリソースの共有があります。そのため、クラスタには、ファイルを一貫性のある方法で共有できるファイルシステムが欠かせません。Sun Cluster ファイルシステムでは、遠隔またはローカルの UNIX 標準 API を使って、ユーザーやアプリケーションからクラスタのどのノードにあるファイルにでもアクセスできます。Sun Cluster システムは、次のファイルシステムをサポートします。

- UNIX ファイルシステム (UFS)
- Sun StorEdge QFS ファイルシステム
- VERITAS ファイルシステム (VxFS)

アプリケーションが、あるノードから別のノードに移動されても、そのアプリケーションは変更なしで同じファイルにアクセスできます。さらに、既存のアプリケーションでクラスタファイルシステムを使用する場合、アプリケーションを変更する必要はありません。

構内クラスタ

標準の Sun Cluster システムは、高可用性と信頼性を 1 箇所から集中的に実現します。地震、洪水、停電などの予測不可能な災害の発生後でもアプリケーションを使用可能なまま維持する必要がある場合は、クラスタを構内クラスタとして構成できます。

構内クラスタでは、数キロメートル 離れた別の建物にノードや共有記憶装置などのクラスタコンポーネントを配置できます。企業構内やその他の場所で、ノードと共有記憶装置を分離し、数キロメートルの範囲にある別の施設内にそれらを配置することが可能です。1 箇所に災害が発生しても、残存するノードが故障したノードのサービスを引き継ぐことができます。これにより、ユーザーは引き続きアプリケーションとデータを使用できます。

障害の監視

Sun Cluster システムでは、多重ホストディスクやマルチパス、グローバルファイルシステムを使って、ユーザーとデータ間のパスの高い可用性を維持します。Sun Cluster システムは、次のコンポーネントの障害を監視します。

- アプリケーション – ほとんどの Sun Cluster データサービスは、データサービスの健全性を周期的に検証するフォルトモニターを備えています。フォルトモニターは、アプリケーションデーモンが動作しているかどうかや、クライアントにサービスが提供されているかどうかを検証します。さらに、フォルトモニターは、検証機能から返される情報に基づいて、デーモンの再起動やフェイルオーバーの指示など、事前に定義されたアクションを開始できます。
- ディスクパス – Sun Cluster ソフトウェアは、ディスクパス監視機能 (DPM) をサポートします。DPM は二次ディスクパスの障害を報告することによって、フェイルオーバーやスイッチオーバーの信頼性を全体的に向上します。
- インターネットプロトコル (IP) マルチパス – Sun Cluster システムで動作する Solaris IP ネットワークマルチパス (IPMP) ソフトウェアは、パブリックネットワークアダプタを監視する基本的なメカニズムです。さらに、障害が検知されると、IPMPソフトウェアは、IP アドレスをあるアダプタから別のアダプタにフェイルオーバーします。

管理と構成のためのツール

Sun Cluster システムのインストールや構成、管理は、SunPlex Manager GUI から行うこともできますし、コマンド行インタフェース (CLI) を使って行うこともできます。

さらに、Sun Cluster システムには、Sun Management Center ソフトウェアの中で動作するモジュールが含まれています。これは、クラスタの一部の作業を行う時の GUI になります。

SunPlex Manager

SunPlex Manager は、Sun Cluster システムの管理に使用するブラウザベースのツールです。管理者は、SunPlex Manager ソフトウェアを通して、システムの管理や監視、ソフトウェアのインストール、システムの構成を行うことができます。

SunPlex Manager ソフトウェアには、次の機能があります。

- 組み込まれたセキュリティーや認証のメカニズム
- Secure Sockets Layer (SSL) のサポート
- 役割によるアクセス制御 (RBAC)
- PAM (Pluggable Authentication Module)
- NAFO (Network Adapter Fail Over) および IPMP グループ管理機能
- クォーラムデバイスやトランスポート、共有ストレージデバイス、リソースグループの管理
- プライベートインターコネクトの高度なエラーチェックや自動検知

コマンド行インタフェース

Sun Cluster コマンド行インタフェースは、Sun Cluster システムのインストールや管理を行ったり、Sun Cluster ソフトウェアのボリュームマネージャー部分を管理する一連のユーティリティーです。

Sun Cluster CLI では、次の SunPlex 管理タスクを行うことができます。

- Sun Cluster 構成の確認
- Sun Cluster ソフトウェアのインストールと構成
- Sun Cluster 構成の更新
- リソースタイプの登録、リソースグループの作成、リソースグループ内のリソースの起動を管理する
- リソースグループとディスクデバイスグループのノードのマスターや状態を変更する
- 役割によるアクセス制御 (RBAC) に基づくアクセス制御
- クラスタ全体の停止

Sun Management Center

Sun Cluster システムには、Sun Management Center ソフトウェアの中で動作するモジュールが含まれています。Sun Management Center ソフトウェアは、管理や監視の操作を行う際のクラスタの基盤となるものです。システム管理者は、GUI や CLI を通して次の作業を行うことができます。

- 遠隔システムの構成
- 性能の監視
- ハードウェアやソフトウェア障害の検知と分離

Sun Management Center ソフトウェアは、Sun Cluster サーバー内での動的再構成 (DR) を管理するインタフェースとしても使用されます。動的再構成には、ドメインの作成や、ボードの動的な接続、動的な切り離しがあります。

役割によるアクセス制御 (RBAC)

従来の UNIX システムでは、root ユーザー (スーパーユーザー) はすべての権限を持ちます。つまり、任意のファイルに対する読み取り権と書き込み権、すべてのプログラムの実行権、および任意のプロセスに終了シグナルを送信する権限があります。Solaris の役割によるアクセス制御 (RBAC) は、スーパーユーザーモデルの権限をすべて与えるか、またはまったく与えないかの機能を置き換えます。RBAC では、基本的に最小限の特権以外は許可しません。つまり、そのユーザーに必要な特権だけを許可します。

RBAC を使用すれば、スーパーユーザーの権限を分割し、それらの権限を特別なユーザーアカウントや役割としてパッケージ化し、それによって、権限を特定の個人に割り当てることができます。このような分割やパッケージ化によって、さまざまなセキュリティーポリシーの作成が可能になります。たとえば、セキュリティーやネットワーキング、ファイアウォール、バックアップ、システム操作など、さまざまな分野で特定目的の管理者用アカウントを設定できます。

第 2 章

Sun Cluster の主要な概念

この章では、Sun Cluster システムのハードウェアやソフトウェアに関連する主な概念を説明します。ユーザーは、Sun Cluster システムを使用する前にこれらの概念を理解しておく必要があります。

この章には、以下の節があります。

クラスタノード

クラスタノードとは、Solaris ソフトウェアと Sun Cluster ソフトウェアが共に動作しているマシンのことです。Sun Cluster ソフトウェアは、2 から 8 ノードのクラスタをサポートします。

クラスタノードは通常、1 つ以上のディスクに接続されます。ディスクに接続されていないノードは、クラスタファイルシステムを使用して多重ホストディスクにアクセスします。PDB データベース構成の下にある各ノードは、一部またはすべてのディスクに同時にアクセスします。

クラスタ内のすべてのノードは、別のノードがいつクラスタに結合されたか、またはクラスタから切り離されたかを認識します。さらに、クラスタ内のすべてのノードは、ローカルに実行されているリソースだけでなく、他のクラスタノードで実行されているリソースも認識します。

同じクラスタ内の各ノードの処理、メモリー、および入出力機能が同等で、パフォーマンスを著しく低下させることなく処理を継続できることを確認してください。フェイルオーバーの可能性があるため、各ノードには、ノードの障害時にもサービスレベル合意を満たせる十分な容量を持つ必要があります。

クラスタインターコネクト

クラスタインターコネクトは、クラスタノード間でクラスタプライベート通信やデータサービス通信を伝送する物理的なデバイス構成です。

冗長なインターコネクトの 1 つに障害が発生しても、操作は残りのインターコネクトを使って続けられます。そのため、システム管理者は、その間に障害を分離し、通信を修復することができます。Sun Cluster ソフトウェアは障害を検知し、修復し、修復されたインターコネクト経由の通信を自動的に再始動します。

詳細については、35 ページの「クラスタインターコネクト」を参照してください。

クラスタメンバーシップ

クラスタメンバーシップモニター (CMM) は、クラスタインターコネクトを使ってメッセージを交換し、次の処理を行う一連の分散エージェントです。

- すべてのノード (定足数) で一貫したメンバーシップの表示を行います。
- メンバーシップの変更に応じて同期のとれた再構成を行います。
- クラスタのパーティション分割を処理します。
- 障害のあるノードを、それが修復されるまでクラスタから除外することによって、すべてのクラスタメンバー間の完全な接続を維持します。

CMM の主な機能はクラスタメンバーシップを確立することですが、そのためには、クラスタに逐次参加するノード群に関してクラスタ全体が合意していなければなりません。CMM は、1 つまたは複数のノード間での通信の途絶など、各ノードにおけるクラスタステータスの大きな変化を検知します。CMM は、トランスポートカーネルモジュールを使ってハートビートを生成し、トランスポート媒体を通してそれをクラスタのほかのノードに伝送します。定義されたタイムアウト時間内にノードからハートビートが送られてこないと、CMM は、そのノードに障害が発生したものとみなし、クラスタの再構成を通してクラスタメンバーシップの再設定を試みます。

CMM は、クラスタメンバーシップを確定し、データの整合性を確保するために、次の処理を行います。

- クラスタへのノードの参加、またはクラスタからのノードの脱退、離脱など、クラスタメンバーシップの変更を考慮します。
- 異常のあるノードを、クラスタから切り離された状態に保ちます。
- 異常のあるノードを、それが修復されるまで非アクティブの状態に保ちます。
- クラスタそのものがノードのサブセットに分割されないように防止します。

クラスタが複数の独立したクラスタに分割されないように防止する方法については、21 ページの「データの完全性」を参照してください。

クラスタ構成レポジトリ

クラスタ構成レポジトリ (CCR) は、クラスタの構成や状態に関する情報を格納するための、クラスタ全体に有効なプライベート分散データベースです。構成データを破損しないために、個々のノードは、クラスタリソースの現在の状態を知っている必要があります。この CCR のおかげで、すべてのノードが、一貫性のあるクラスタ像を持つことができます。CCR は、エラーや復旧の情況が発生したり、クラスタの一般的なステータスに変化があると更新されます。

CCR 構造には、次のような情報が含まれています。

- クラスタ名とノード名
- クラスタトランスポート構成
- Solaris ボリュームマネージャーディスクセットや VERITAS ディスクグループの名前
- 個々のディスクグループをマスターできるノードのリスト
- データサービスの操作に関するパラメータ値
- データサービスコールバックメソッドへのパス
- DID デバイス構成
- クラスタの現在のステータス

フォルトモニター

Sun Cluster システムでは、アプリケーションそのものや、ファイルシステム、ネットワークインタフェースを監視することによって、ユーザーとデータ間の「パス」にあるすべてのコンポーネントの高い可用性を保ちます。

Sun Cluster ソフトウェアは、ノードを素早く検知し、そのノードと同等のリソースを備えたサーバーを作成します。Sun Cluster ソフトウェアのおかげで、障害のあるノードの影響を受けないリソースはこの復旧中も引き続き使用され、障害のあるノードのリソースは復旧すると同時に再び使用可能になります。

データサービス監視

Sun Cluster の各データサービスには、データサービスを定期的に検査してその状態を判断するフォルトモニターがあります。フォルトモニターは、アプリケーションデーモンが動作しているかどうかや、クライアントにサービスが提供されているかどうかを検証します。探索によって得られた情報をもとに、デーモンの再起動やフェイルオーバーの実行などの事前に定義された処置が開始されます。

ディスクパスの監視

Sun Cluster ソフトウェアは、ディスクパス監視 (DPM) がサポートします。DPM は、二次ディスクパスの障害を報告することによって、フェイルオーバーやスイッチオーバーの全体的な信頼性を高めます。ディスクパスの監視には 2 つの方法があります。1 つめの方法は `scdpm` コマンドを使用する方法です。このコマンドを使用すると、クラスタ内のディスクパスの状態を監視、監視解除、または表示できます。コマンド行オプションの詳細については、`scdpm(1M)` のマニュアルページを参照してください。

2 つめの方法は、SunPlex Manager の GUI (Graphical User Interface) を使用してクラスタ内のディスクパスを監視する方法です。SunPlex Manager では、監視されているディスクパスがトポロジで表示されます。このトポロジビューは 10 分ごとに更新され、失敗した ping の数が表示されます。

IP マルチパス監視

各クラスタノードには独自の IPMP 構成があり、これは他のクラスタノード上の構成と異なる場合があります。IP ネットワークマルチパス は、次のネットワークの通信障害を監視します。

- ネットワークアダプタの送信/受信パスがパケットの伝送を停止した。
- ネットワークアダプタとリンクとの接続がダウンしている。
- Ethernet スイッチ上のポートがパケットを送受信しない。
- グループ内の物理インタフェースがシステムの起動時に存在しない。

定足数デバイス

クォーラムデバイスとは、定足数 (quorum) を確立してクラスタを実行するために使用される「票」を持つ、複数のノードによって共有されるディスクです。クラスタは、票の定足数が満たされた場合にのみ動作可能です。クォーラムデバイスは、クラスタが独立したノードの集合にパーティション分割されたときに、どちらのノード集合が新しいクラスタを構成するかを確定するために使用されます。

クラスタノードと定足数デバイスはどちらも、定足数を確立するために投票します。デフォルトにより、クラスタノードは、起動してクラスタメンバーになると、定足数投票数 (quorum vote count) を 1 つ獲得します。ノードは、ノードのインストール中や管理者がノードを保守状態にした時には、投票数は 0 になります。

クォーラムデバイスは、デバイスへのノード接続の数に基づいて投票数を獲得します。クォーラムデバイスは、設定されると、最大投票数 N-1 を獲得します。この場合、N は、投票数がゼロ以外で、クォーラムデバイスへ接続された投票数を示します。たとえば、2 つのノードに接続された、投票数がゼロ以外のクォーラムデバイスの投票数は 1 (2–1) になります。

データの完全性

Sun Cluster システムはデータ破損を防ぎ、データの完全性を保とうとします。それぞれのクラスタノードはデータとリソースを共有していますので、クラスタが、同時にアクティブである複数のパーティションに分割されることがあってはなりません。CMM は、必ず 1 つのクラスタだけが使用可能であることを保証します。

クラスタのパーティション分割によって起こる問題に split-brain と amnesia があります。split-brain が起こるのは、ノード間のクラスタインターコネクトが失われ、クラスタがサブクラスタにパーティション分割され、各サブクラスタが唯一のパーティションであると認識する場合です。ほかのサブクラスタの存在を認識していないサブクラスタは、ネットワークアドレスの重複やデータ破損など、共有リソースの対立を引き起こすおそれがあります。

amnesia は、すべてのノードがそのクラスタ内で不安定なグループの状態になっている場合に起こります。たとえば、ノード A とノード B からなる 2 ノードクラスタがあるとします。ノード A が停止すると、CCR の構成データはノード B のものだけが更新され、ノード A のものは更新されません。この後でノード B が停止し、ノード A が再起動されると、ノード A は CCR の古い内容に基づいて動作することになります。この状態を amnesia と呼びます。この状態になると、クラスタは、古い構成情報で実行されることがあります。

split-brain と amnesia の問題は、各ノードに 1 票を与え、過半数の投票がないとクラスタが動作しないようにすることで防止できます。過半数の投票を得たパーティションは「定足数 (quorum)」を獲得し、アクティブになります。この過半数の投票メカニズムは、クラスタのノード数が 2 を超える場合には有効です。しかし、2 ノードクラスタでは過半数が 2 であるため、このようなクラスタがパーティション分割されると、パーティションは外部からの投票で定足数を獲得します。この外部からの投票は、クォーラムデバイスによって行われます。定足数デバイスは、2 つのノードで共有されている任意のディスクにすることができます。

表 2–1に、 Sun Cluster ソフトウェアが定足数を使用して split-brain と anmesia を回避する様子を示します。

表 **2–1** クラスタ定足数、および split-brain と amnesia の問題

問題	定足数による解決策
split brain	過半数の投票を獲得したパーティション (サブクラスタ) だけをクラスタとして実行できるようにする (過半数を獲得できるパーティションは 1 つのみ)。ノードが定足数を獲得できないと、ノードはパニックになります。
amnesia	起動されたクラスタには、最新のクラスタメンバーシップのメンバーであった (したがって、最新の構成データを持つ) ノードが少なくとも 1 つあることを保証する。

障害による影響の防止

クラスタの主要な問題は、クラスタがパーティション分割される (sprit-brain と呼ばれる) 原因となる障害です。このような状態になると、一部のノードが通信できなくなるため、個々のノードまたはノードの一部が、ノード単体または一部のノードによってクラスタを形成しようとします。各部分、つまりパーティションは、多重ホストディスクに対して単独のアクセスと所有権を持つものと誤って認識します。しかし、複数のノードがディスクに書き込もうとすると、データ破損を招くおそれがあります。

二重障害の防止機能は、ディスクへのアクセスを防止することによってノードが多重ホストディスクにアクセスすることを制限します。障害が発生するかパーティション分割され、ノードがクラスタから切り離されると、障害による影響の防止機能によって、ノードがディスクにアクセスできなくなります。現在のメンバーノードだけが、ディスクへのアクセス権を持つため、データの完全性が保たれます。

Sun Cluster システムは、SCSI ディスクリザベーションを使用して、二重障害の防止機能を実装します。SCSI 予約を使用すると、障害が発生したノードは、多重ホストディスクによって阻止されて、これらのディスクへのアクセスが防止されます。

クラスタメンバーは、別のノードがクラスタインターコネクトを介して通信していないことを検出すると、二重障害の防止手順を開始して、障害のあるそのノードが共有ディスクへアクセスするのを防止します。この二重障害の防止機能が動作すると、アクセスを阻止されたノードはパニック状態になり、そのコンソールに「reservation conflict」メッセージが表示されます。

障害の影響を防止するフェイルファースト機構

フェイルファースト機構は、障害のノードをパニック状態にしますが、そのノードの再起動を防ぐことはしません。パニックの後にこのノードは再起動を行ない、クラスタに再び参加しようとすることがあります。

定足数を獲得できるパーティションに属していないノードが、クラスタ内の他のノードとの接続を失うと、そのノードは別のノードによってクラスタから強制的に切り離されます。定足数を獲得できるパーティションに属しているノードはすべて、共有ディスク上でリザベーションを発行します。フェイルファースト機構の結果、定足数を満たしていないノードはパニック状態になります。

デバイス

グローバルファイルシステムでは、クラスタのすべてのファイルがすべてのノードから同じように認識され、アクセス可能になります。それと同様に、Sun Cluster ソフトウェアの下では、クラスタのすべてのデバイスがクラスタ全体から認識され、アクセス可能になります。つまり、どのノードからでも入出力サブシステムを通してクラスタのどのデバイスにもアクセスできます。デバイスが物理的にどこに接続されているかは関係ありません。このアクセスをグローバルデバイスアクセスと呼びます。

グローバルデバイス

Sun Cluster システムでは、クラスタの任意のデバイスに任意のノードから高い可用性をもってクラスタレベルでアクセスできるようにするために、グローバルデバイスを使用します。通常、ノードからグローバルデバイスにアクセスできないことがあると、Sun Cluster ソフトウェアは、そのデバイスへのパスを別のパスに切り替え、アクセスをそのパスに振り向けます。グローバルデバイスでは、この変更は簡単です。どのパスを使用する場合でも、デバイスには同じ名前が使用されるからです。リモートデバイスへのアクセスは、同じ名前を持つローカルデバイスの場合と同じように行われます。さらに、クラスタのグローバルデバイスにアクセスする API は、ローカルのデバイスにアクセスする API と同じです。

Sun Cluster グローバルデバイスには、ディスク、CD-ROM、テープが含まれます。ただし、サポートされるマルチポートのグローバルデバイスはディスクだけです。つまり、CD-ROM とテープは、現在可用性の高いデバイスではありません。各サーバーのローカルディスクも多重ポート化されていないため、可用性の高いデバイスではありません。

クラスタは、クラスタ内の各ディスク、CD-ROM、テープデバイスに一意の ID を割り当てます。この割り当てによって、クラスタ内の任意のノードから各デバイスに対して一貫したアクセスが可能になります。

デバイス ID

Sun Cluster ソフトウェアは、デバイス ID (DID) ドライバと呼ばれるコンストラクトを通してグローバルデバイスを管理します。このドライバを使用して、多重ホストディスク、テープドライブ、CD-ROM を含め、クラスタ内のあらゆるデバイスに一意の ID を自動的に割り当てます。

DID ドライバは、クラスタのグローバルデバイスアクセス機能の重要な部分です。DID ドライバは、クラスタのすべてのノードを検査し、一意のディスクデバイスからなるリストを構築します。さらに、DID ドライバは、一意のメジャー番号とマイナー番号を各デバイスに割り当てます。この数字は、クラスタのすべてのノードで一貫性をもって管理されます。グローバルデバイスへのアクセスは、従来の Solaris DID と替わって DID ドライバによって割り当てられた 一意の DID を使って行われます。

このような方法をとれば、Solaris ボリュームマネージャーや Sun Java System Directory Server など、ディスクにアクセスするアプリケーションが何であれ、クラスタ全体で一貫性のあるパスが使用されます。多重ホストディスクの場合は、この一貫性がとりわけ重要です。各デバイスのローカルのメジャー番号とマイナー番号はノードによって異なる可能性があるからです。さらに、これらの数字は、Solaris デバイスの命名規約も同様に変更する可能性があります。

ローカルデバイス

Sun Cluster ソフトウェアはローカルデバイスも管理します。このようなデバイスは、サービスが動作しているノードだけでアクセスされるものであり、クラスタに物理的に接続されています。ローカルデバイスは、性能の点でグローバルデバイスよりも有利です。ローカルデバイスでは、状態情報を複数のノードに同時に複製する必要がないからです。デバイスのドメインに障害が発生すると、そのデバイスにはアクセスできなくなります。ただし、そのデバイスを複数のノードで共有できる場合を除きます。

ディスクデバイスグループ

ディスクデバイスグループでは、ボリュームマネージャーのディスクグループが「グローバル」になります。ボリュームマネージャーは、使用しているディスクに対してマルチパスと多重ホストをサポートするからです。多重ホストディスクに物理的に接続された各クラスタノードは、ディスクデバイスグループへのパスを提供します。

Sun Cluster システムで Sun Cluster ソフトウェアを使用している多重ホストディスクを制御するには、多重ホストディスクをディスクデバイスグループとして登録します。この登録によって、Sun Cluster システムは、どのノードがどのボリュームマネージャーディスクグループへのパスをもっているかを知ることができます。Sun Cluster ソフトウェアは、クラスタ内のディスクデバイスやテープデバイスごとに、raw ディスクデバイスグループを作成します。これらのクラスタデバイスグループは、ユーザーがグローバルファイルシステムをマウントするか、raw データベースファイルにアクセスすることによって、これらのデバイスグループをグローバルデバイスとしてアクセスするまでオフライン状態に置かれます。

データサービス

データサービスは、Sun Cluster 構成の下でアプリケーションを変更なしで実行できるようにする、ソフトウェアと構成ファイルの組み合わせです。Sun Cluster 構成の下で動作するアプリケーションは、リソースグループマネージャー (RGM) の制御下にある 1 つのリソースです。データサービスを使えば、Sun Java System Web Server や Oracle データベースなどのアプリケーションをクラスタで (単一のサーバーではなく) 実行するように構成できます。

データサービスのソフトウェアは、アプリケーションに対して次の操作を行う Sun Cluster 管理メソッドを実装しています。

- アプリケーションの起動
- アプリケーションの停止
- アプリケーションの障害の監視とこの障害からの復旧

データサービスの構成ファイルは、RGM にとってアプリケーションを意味するリソースのプロパティーを定義したものです。

クラスタのフェイルオーバーデータサービスやスケーラブルデータサービスの処理は RGM によって制御されます。RGM は、クラスタメンバーシップの変更に応じて選択されたクラスタのノードでデータサービスの起動や停止を行います。データサービスアプリケーションは、RGM を通してクラスタフレームワークを利用できます。

RGM はデータサービスをリソースとして制御します。これらの実装は Sun によって提供されるか、開発者によって作成されます。後者の場合には、汎用的なデータサービステンプレートや、 データサービス開発ライブラリ API (DSDL API)、リソース管理 API (RMAPI) が使用されます。クラスタ管理者は、リソースグループと呼ばれる入れ物 (コンテナ) の中でリソースの作成や管理を行います。リソースやリソースグループの状態は、RGM や管理者のアクションによってオンラインやオフラインにされます。

リソースタイプ

リソースタイプとは、あるアプリケーションをクラスタに説明するプロパティーの集まりのことです。この集合には、クラスタのノードでアプリケーションをどのように起動、停止、監視するかを示す情報が含まれています。さらに、リソースタイプには、アプリケーションをクラスタで使用するために必要なアプリケーション固有のプロパティーも含まれています。Sun Cluster データサービスには、いくつかのリソースタイプが事前に定義されています。たとえば、Sun Cluster HA for Oracle のリソースタイプは `SUNW.oracle-server`、Sun Cluster HA for Apache のリソースタイプ `SUNW.apache` です。

リソース

リソースとは、クラスタ規模で定義したリソースタイプのインスタンスのことです。リソースタイプを使用すると、アプリケーションの複数のインスタンスをクラスタにインストールできます。ユーザーがリソースを初期化すると、RGM は、アプリケーション固有のプロパティーに値を割り当てます。リソースは、リソースタイプのレベルにあるすべてのプロパティーを継承します。

データサービスは、いくつかのタイプのリソースを使用します。たとえば、Apache Web Server や Sun Java System Web Server などのアプリケーションは、それらが依存するネットワークアドレス (論理ホスト名と共有アドレス) を使用します。アプリケーションとネットワークリソースは RGM が管理する基本単位です。

リソースグループ

RGM は、複数のリソースをリソースグループという 1 つの単位として扱うことができるようにします。リソースグループとは、関連する (あるいは、相互に依存する) リソースの集合のことです。たとえば、`SUNW.LogicalHostname` リソースタイプから派生したリソースは、Oracle データベースリソースタイプから派生したリソースと同じリソースグループに置かれることがあります。リソースグループ上でフェイルオーバーまたはスイッチオーバーが開始されると、リソースグループは 1 つの単位として移行されます。

データサービスのタイプ

データサービスを使用すると、アプリケーションは可用性の高いものやスケーラブルなサービスになります。クラスタで単一の障害が発生した場合、大幅なアプリケーションの中断を回避できます。

データサービスを構成する際には、次のデータサービスのタイプから 1 つを選択する必要があります。

- フェイルオーバーデータサービス
- スケーラブルデータサービス
- パラレルデータサービス

フェイルオーバーデータサービス

フェイルオーバーとは、クラスタがアプリケーションを障害のある稼動系から、指定の冗長化された待機系に自動的に再配置するプロセスのことをいいます。フェイルオーバーアプリケーションには、次の特徴があります。

- クラスタの 1 つのノードだけに実行の資格があります。
- クラスタで動作していることを意識させません。
- クラスタフレームワークに基づいて HA を達成します。

フォルトモニターは、エラーを検出すると、データサービスの構成に従って、同じノードでそのインスタンスを再起動しようとするか、別のノードでそのインスタンスを起動 (フェイルオーバー) しようとします。フェイルオーバーサービスは、アプリケーションインスタンスリソースとネットワークリソース (論理ホスト名) のコンテナである、フェイルオーバーリソースグループを使用します。論理ホスト名とは、1 つのノードに構成して、後で自動的に元のノードや別のノードに構成できる IP アドレスのことです。

サービスが一時的に中断されるため、クライアントは、フェイルオーバーの完了後にサービスに再接続しなければならない場合があります。しかし、クライアントは、サービスの提供元である物理サーバーが変更したことを意識しません。

スケーラブルデータサービス

スケーラブルデータサービスでは、複数のアプリケーションインスタンスが複数のノードで同時に動作します。スケーラブルサービスは、2 つのリソースグループを使用します。スケーラブルリソースグループにはアプリケーションリソースが、フェイルオーバーリソースグループには、スケーラブルサービスが依存するネットワークリソース (共有アドレス) がそれぞれ含まれています。スケーラブルリソースグループは、複数のノードでオンラインにできるため、サービスの複数のインスタンスを同時に実行できます。共有アドレスのホストとなるフェイルオーバーリソースグループは、一度に 1 つのノードでしかオンラインにできません。スケーラブルサービスをホストとするすべてのノードは、サービスをホストするための同じ共有アドレスを使用します。

クラスタは、同一のネットワークインタフェース (グローバルインタフェース) を通してサービス要求を受け取ります。これらの要求は、事前に定義されたいくつかのアルゴリズムの 1 つに基づいてノードに分配されます (アルゴリズムは負荷均衡ポリシーによって設定される)。クラスタは、負荷均衡ポリシーを使用し、いくつかのノード間でサービス負荷均衡をとることができます。

パラレルアプリケーション

Sun Cluster システムは、パラレルデータベース (PDB) を使用することによってクラスタのすべてのノードでアプリケーションを並列で実行できるようにする環境を提供します。Sun Cluster Support for Oracle Parallel Server/Real Application Clusters は、Oracle Parallel Server/Real Application Clusters を Sun Cluster ノードで実行できるようにするパッケージ群です。さらに、このデータサービスでは、Sun Cluster コマンドを使って Sun Cluster Support for Oracle Parallel Server/Real Application Clusters を管理できます。

パラレルアプリケーションはクラスタ環境で動作するように考えられたものです。したがって、このようなアプリケーションは、複数のノードから同時マスターされます。Oracle Parallel Server/Real Application Clusters 環境では、複数の Oracle インスタンスが協力して同じ共有データベースにアクセスします。Oracle クライアントは、任意のインスタンスを使用してデータベースにアクセスできます。したがって、1 つまたは複数のインスタンスで障害が発生しても、クライアントは残りのインスタンスに接続することによって、引き続きデータベースにアクセスできます。

第 3 章

Sun Cluster のアーキテクチャー

Sun Cluster アーキテクチャーでは、一連のシステムが単一の大規模システムとして配備され、管理され、認識されます。

この章には、以下の節があります。

Sun Cluster のハードウェア環境

クラスタは、次のハードウェアコンポーネントから構成されます。

- ローカルディスク (非共有) を備えたクラスタノード。クラスタの主要なコンピューティングプラットフォームです。
- 多重ホストストレージ。ノード間で共有されるディスクです。
- テープや CD-ROM などのリムーバブルメディア。グローバルデバイスとして構成されます。
- クラスタインターコネクト。ノード間の通信チャネルとして使用されます。
- パブリックネットワークインタフェース。クライアントシステムによって使用されるネットワークインタフェースは、このインタフェースを通してクラスタのデータサービスにアクセスします。

図 3–1に、ハードウェアコンポーネント相互間の連携のしくみを示します。

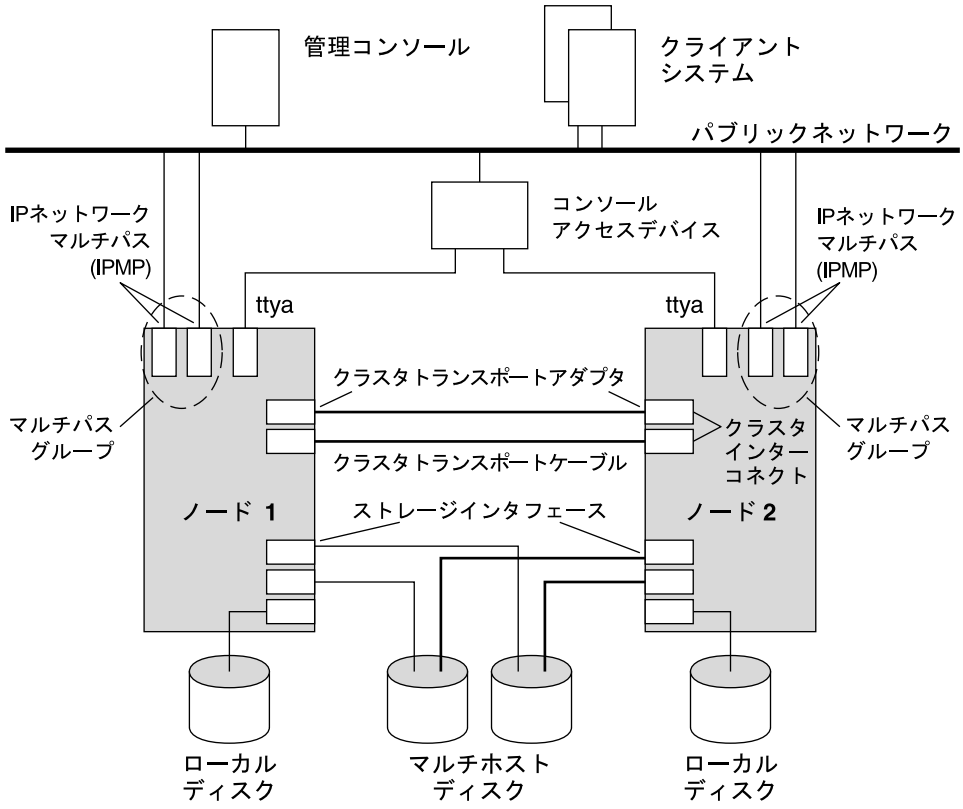

図 **3–1** Sun Cluster ハードウェアコンポーネント

Sun Cluster のソフトウェア環境

ノードがクラスタメンバーとして動作するためには、ノードに次のソフトウェアがインストールされていなければなりません。

- Solaris ソフトウェア
- Sun Cluster ソフトウェア
- データサービスアプリケーション
- ボリューム管理 (Solaris™ Volume Manager または VERITAS Volume Manager)

  ただし、そのボックス自体のボリューム管理を使用する構成は例外です。この構成では、ソフトウェアボリュームマネージャーが必要ない場合があります。

図 3–2に、相互に機能して Sun Cluster ソフトウェア環境を構成するソフトウェアコンポーネントの概要を示します。

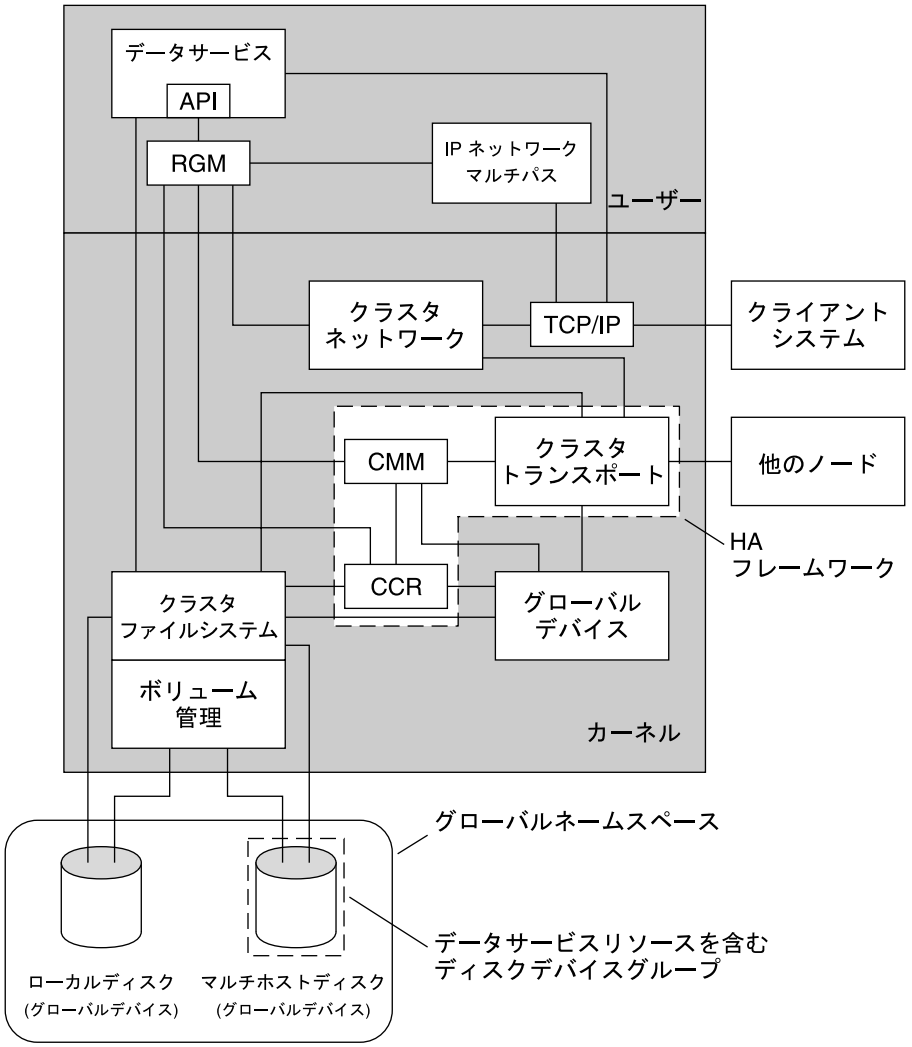

図 **3–2** Sun Cluster ソフトウェアアーキテクチャー

クラスタメンバーシップモニター

データが破壊から保護されるように保証するには、すべてのノードが、クラスタメンバーシップに対して一定の同意に達していなければなりません。必要であれば、CMM は、障害に応じてクラスタサービスのクラスタ再構成を調整します。

CMM は、クラスタのトランスポート層から、他のノードへの接続に関する情報を受け取ります。CMM は、クラスタインターコネクトを使用して、再構成中に状態情報を交換します。

CMM は、クラスタメンバーシップの変更を検出すると、それに合わせてクラスタを構成します。この構成処理では、クラスタリソースが、クラスタの新しいメンバーシップに基づいて再配布されることがあります。

CMM は完全にカーネル内で動作します。

クラスタ構成レポジトリ (CCR)

CCR は、CMM に依存して、定足数 (quorum) が確立された場合にのみクラスタが実行されるように保証します。CCR は、クラスタ全体のデータの一貫性を確認し、必要に応じて回復を実行し、データへの更新を容易にします。

クラスタファイルシステム

クラスタファイルシステムは、次のコンポーネント間のプロクシです。

- あるノード上のカーネルとそのノードが使用しているファイルシステム
- そのディスク (1 つまたは複数) と物理的に接続されているノード上のボリュームマネージャー

クラスタファイルシステムでは、グローバルデバイス (ディスク、テープ、CD-ROM) が使用されます。グローバルデバイスには、クラスタのどのノードからでも同じファイル名 (たとえば、`/dev/global/`) を使ってアクセスできます。そのノードは、アクセスするストレージデバイスに物理的に接続されている必要はありません。ユーザーは、グローバルデバイスを通常のデバイスと同じように使用できます。つまり、`newfs` や `mkfs` を使ってグローバルデバイスにファイルシステムを作成することができます。

クラスタファイルシステムには、次の機能があります。

- ファイルのアクセス場所が透過的になります。システムのどこにあるファイルでも、プロセスから開くことができます。さらに、すべてのノードのプロセスから同じパス名を使ってファイルにアクセスできます。

  注 – クラスタファイルシステムは、ファイルを読み取る際に、ファイル上のアクセス時刻を更新しません。

- 一貫したプロトコルを使用して、ファイルが複数のノードから同時にアクセスされている場合でも、UNIX ファイルアクセス方式を維持します。

- 拡張キャッシュ機能とゼロコピーバルク入出力移動機能により、ファイルデータを効率的に移動することができます。
- クラスタファイルシステムには、`fcntl(2)` インタフェースに基づく、高度な可用性を備えたアドバイザリファイルロッキング機能があります。クラスタファイルシステムのファイルに対してアドバイザリファイルロッキング機能を使えば、複数のクラスタノードで動作するアプリケーションの間で、データのアクセスを同期化できます。ファイルロックを所有するノードがクラスタから切り離されたり、ファイルロックを所有するアプリケーションが異常停止すると、それらのロックはただちに解放されます。
- 障害が発生した場合でも、データへの連続したアクセスが可能です。アプリケーションは、ディスクへのパスが有効であれば、障害による影響を受けません。この保証は、raw ディスクアクセスとすべてのファイルシステム操作で維持されます。
- クラスタファイルシステムは、基本のファイルシステムからもボリュームマネージャーからも独立しています。クラスタシステムファイルは、サポートされているディスク上のファイルシステムすべてを広域にします。

スケーラブルデータサービス

クラスタネットワーキングの主な目的は、データサービスにスケーラビリティーを提供することにあります。スケーラビリティーとは、サービスに提供される負荷が増えたときに、新しいノードがクラスタに追加されて新しいサーバーインスタンスが実行されるために、データサービスがこの増加した負荷に対して一定の応答時間を維持できるということを示します。スケーラブルデータサービスの例としては、Web サービスがあります。通常、スケーラブルデータサービスはいくつかのインスタンスからなり、それぞれがクラスタの異なるノードで実行されます。これらのインスタンスは、遠隔クライアントに対して単一のサービスとして動作し、そのサービスの機能を提供します。別々のノードで動作するいくつかの `httpd` デーモンからなるスケーラブル Web サービスでは、任意のデーモンでクラスタ要求を処理できます。要求に対応するデーモンは、負荷均衡ポリシーによって決められます。クライアントへの応答は、その要求にサービスを提供する特定のデーモンからではなく、サービスからのもののようにみえるため、単一サービスの外観が維持されます。

次の図は、スケーラブルサービスの構造を示したものです。

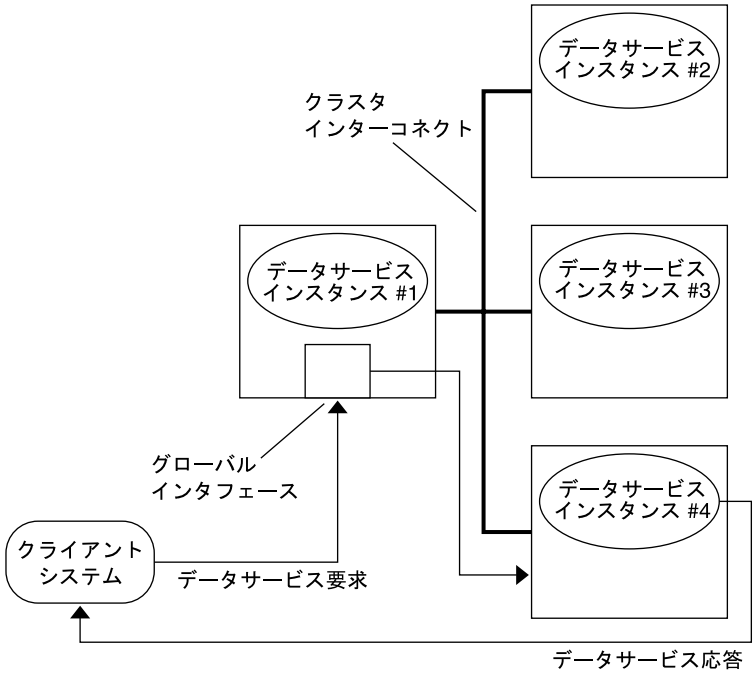

図 **3–3** スケーラブルデータサービスの構造

グローバルインタフェースのホストではないノード (プロキシノード) には、そのループバックインタフェースでホストされる共有アドレスがあります。グローバルインタフェースに入ってくるパケットは、構成可能な負荷均衡ポリシーに基づいてほかのクラスタノードに分配されます。次に、構成できる負荷均衡ポリシーについて説明します。

負荷均衡ポリシー

負荷均衡は、スケーラブルサービスのパフォーマンスを応答時間とスループットの両方の点で向上させます。

スケーラブルデータサービスには、*pure* と *sticky* の 2 つのクラスがあります。pure サービスとは、どのインスタンスでもクライアント要求に応答できるサービスをいいます。sticky サービスでは、ノードへの要求の負荷をクラスタが均衡させます。これらの要求は、別のインスタンスには変更されません。

pure サービスは、ウェイト設定した (weighted) 負荷均衡ポリシーを使用します。この負荷均衡ポリシーのもとでは、クライアント要求は、デフォルトで、クラスタ内のサーバーインスタンスに一律に分配されます。たとえば、各ノードのウェイトが 1 であるような 3 ノードクラスタでは、各ノードが、任意のクライアントからの要求をそのサービスのために 3 分の 1 ずつ処理します。ウェイトの変更は、`scrgadm(1M)` コマンドインタフェースか SunPlex Manager GUI を使っていつでもできます。

sticky サービスには、*ordinary sticky* と *wildcard sticky* があります。sticky サービスを使用すると、内部状態メモリーを共有でき (アプリケーションセッション状態)、複数の TCP 接続でアプリケーションレベルの同時セッションが可能です。

ordinary sticky サービスを使用すると、クライアントは、複数の同時 TCP 接続で状態を共有できます。このクライアントを、単一ポートで待機するサーバーインスタンスに対して「sticky」であるといいます。クライアントは、インスタンスが起動していてアクセス可能であり、負荷均衡ポリシーがサービスのオンライン時に変更されていなければ、すべての要求が同じサーバーのインスタンスに送られることを保証されます。

wildcard sticky サービスは、動的に割り当てられたポート番号を使用しますが、クライアント要求が同じノードに送りかえされると想定します。クライアントは、同じ IP アドレスに対して、複数のポート間で *sticky wildcard* であるといいます。

多重ホストディスク記憶装置

Sun Cluster ソフトウェアは、複数のノードに同時に接続できる多重ホストディスクストレージを使用することによって、ディスクの高い可用性を実現します。これらのディスクは、ボリューム管理ソフトウェアの使用を通して、クラスタノードからマスターされる共有ストレージに編成されます。そして、障害が発生したときに別のノードに移動されるように構成されます。Sun Cluster システムで多重ホストディスクを使用することには、さまざまな利点があります。たとえば、次はその例です。

- ファイルシステムへのグローバルアクセス
- ファイルシステムやデータへの複数のアクセスパス
- 単一ノード障害に耐えられる

クラスタインターコネクト

少なくとも2 つの冗長な物理的に独立したネットワーク、またはパスを使用して、すべてのノードをクラスタインターコネクトによって接続し、単一障害を回避する必要があります。冗長性を保つためには 2 つのインターコネクトが必要ですが、ボトルネックを解消したり、冗長性や拡張性を強化するために、最大 6 つのインターコネクトを使ってトラフィックを分散させることができます。Sun Cluster インターコネクトでは、Fast Ethernet、InfiniBand、Gigabit-Ethernet、Sun Fire Link、または Scalable Coherent Interface (SCI, IEEE 1596-1992) の使用を通して、高性能のクラスタ内通信がサポートされます。

クラスタ環境のノード間通信には、高速、低遅延のインターコネクトとプロトコルが欠かせません。Sun Cluster システムの SCI インターコネクトは、一般的なネットワークインタフェースカード (NIC) よりも高い性能を発揮します。Sun Cluster で

は、Sun Fire Link ネットワークにおけるノード間通信に Remote Shared Memory (RSM™) インタフェースを使用します。RSM は、非常に効率的な遠隔メモリー操作を行う Sun メッセージングインタフェースです。

RSM Reliable Datagram Transport (RSMRDT) ドライバは、RSM API 上に構築されるドライバと、RSMRDT-API インタフェースをエクスポートするライブラリから構成されます。このドライバは、Oracle Parallel Server/Real Application Clusters の性能を向上させます。このドライバはまた、負荷均衡機能と高可用性 (HA) 機能をドライバ内部で直接提供することにより、両機能を強化すると共に、クライアントからの透過な利用を可能にしています。

クラスタインターコネクトは、以下のハードウェアコンポーネントで構成されます。

- アダプタ – 個々のクラスタノードに存在するネットワークインタフェースカード。複数のインタフェースを持つネットワークアダプタは、アダプタ全体に障害が生じると、単一地点による障害の原因となる可能性があります。
- 接続点 – クラスタノードの外部にあるスイッチ。ジャンクションは、パススルーおよび切り換え機能を実行して、3 つ以上のノードに接続できるようにします。2 ノードクラスタでは、冗長な物理ケーブルによってノードが相互に直接接続されるため、ジャンクションは必要ありません。これらの冗長化なケーブルは、各ノードの冗長化されたアダプタに接続されます。3 ノード以上の構成では、ジャンクションが必要です。
- ケーブル – 2 つのネットワークアダプタまたはアダプタとジャンクションの間をつなぐ物理接続。

図 3–4に、3 つのコンポーネントがどのように接続されているかを示します。

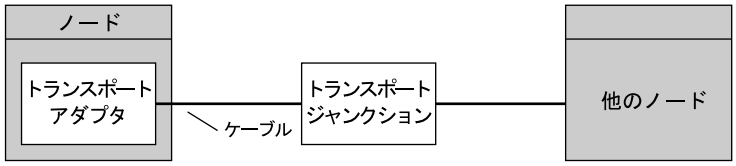

図 **3–4** クラスタインターコネクト

IP ネットワークマルチパスグループ

パブリックネットワークアダプタは、IPMP グループ (マルチパスグループ) として編成されます。各マルチパスグループには、1 つまたは複数のパブリックネットワークアダプタがあります。マルチパスグループの各アダプタはアクティブにすることができます。あるいは、スタンバイインタフェースを構成し、フェイルオーバーが起こるまでそれらを非アクティブにしておくことができます。

マルチパスグループは、論理ホスト名と共有アドレスリソースの基盤です。つまり、ノード上の同じマルチパスグループは、任意の数の論理ホスト名または共有アドレスリソースをホストできます。マルチパスを作成すれば、クラスタノードのパブリックネットワーク接続を監視できます。

論理ホスト名や共有アドレスリソースについては、『Sun Cluster データサービスの計画と管理 (Solaris OS 版)』を参照してください。

パブリックネットワークインタフェース

クライアントは、パブリックネットワークインタフェースを介してクラスタに接続します。各ネットワークアダプタカードは、カードに複数のハードウェアインタフェースがあるかどうかによって、1 つまたは複数のパブリックネットワークに接続できます。複数のパブリックネットワークインタフェースカードをもつノードを設定することによって、複数のカードをアクティブにし、それぞれを相互のフェイルオーバーバックアップとすることができます。アダプタの 1 つに障害が発生すると、Sun Cluster の Solaris IPMP ソフトウェアが呼び出され、障害のあるインタフェースが同じグループの別のアダプタにフェイルオーバーされます。

索引